ML 算法探索

吴天阳 2023 年 10 月 28 日

考虑结合经典 VAE 和Regularizing Deep Networks with Semantic Data Augmentation(ISDA)一个对损失函数加入隐参数分布从而对进行数据增强的方法.

1 Motivation

1. 传统 VAE 只能做到图像特征分布计算和图像重建，数据增强只能简单地对特征加入噪声然后进行重建，无法对某一指定类别进行某一特征方向上的变化.
2. ISDA 需要基于数据集重新计算指定类别的协方差矩阵，并不高效.

目标：基于一个**带有分类和重建任务 VAE 模型**，使其能够在一个带有类别的数据集上实现对**指定类别**的数据生成，要求**端到端**直接训练和生成，无需基于数据集重新计算数据的当前平均分布或协方差.

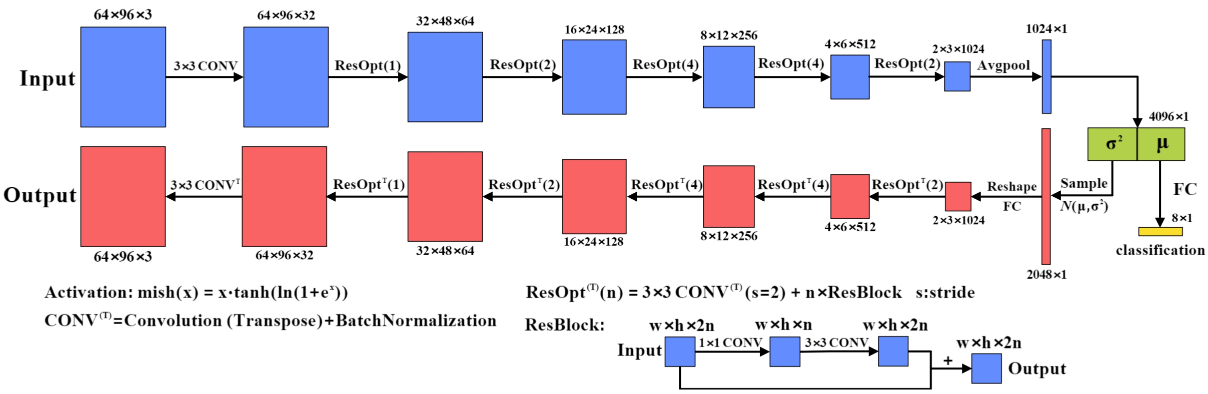


图 1: G-VAE 模型在 Celeba 数据集上的网络框架

2 Method

2.1 VAE 变分自动编码机

设两个可以通过神经网络得到的分布：

- 编码器：图像 x 对应而隐参数分布 $q_\phi(z|x) = \mathcal{N}(\mu_\phi(x), \Sigma_\phi(x))$.
- 解码器：隐参数 z 对应的图像分布 $p_\theta(x|z) = \mathcal{N}(\mu_\theta(x), \Sigma_\theta(x))$.

其中 ϕ, θ 分别为编码器和解码器对应的网络参数.

通过变分方法：

$$\begin{aligned}\log p_\theta(x|z) &= \int q(z)\log p_\theta\theta(x|z)\,\mathrm{d}z = \int q(z)\log\frac{p_\theta(x,z)}{p(z)}\cdot\frac{q(z)}{q(z)}\,\mathrm{d}z\\ &= \mathrm{KL}(q||p) + \int q(z)\log\frac{p_\theta(x,z)}{q(z)}\,\mathrm{d}z\\ &\geqslant \int q_\phi(z)\log\frac{p_\theta(x,z)}{q_\phi(z)}\,\mathrm{d}z \quad (\text{用}q_\phi(z)\to q(z)\text{得到变分下界})\end{aligned}\tag{2.1}$$

于是 $\max_{\theta} \log p_\theta(x|z) \iff \max_{\theta,\phi} \int q_\phi(z) \log \frac{p_\theta(x,z)}{q_\phi(z)}\, \mathrm{d}z$，又由于

$$
\begin{aligned}
&\max_{\theta,\phi} \int q_\phi(z) \log \frac{p_\theta(x,z)}{q_\phi(z)}\, \mathrm{d}z = \int q_\phi(z) \log \frac{p(z)p_\theta(x|z)}{q_\phi(z)}\, \mathrm{d}z \\
\iff &\max_{\theta,\phi} -\mathrm{KL}(q_\phi||p) + \mathbb{E}_{z\sim q_\phi}[\log p_\theta(x|z)] \\
\iff &\min_{\theta,\phi} \mathrm{KL}(q_\phi||p) + ||x-\mu_\theta(z)||_2^2
\end{aligned}
$$

第一项，对解码器参数的更新：设隐空间维度为 K，令 $p(z) \sim \mathcal{N}(0, I_K)$ 作为隐参数的目标分布，假设 $q_\phi(z)$ 独立同分布 $\mathcal{N}(\mu, \mathrm{diag}(\sigma_1^2, \cdots, \sigma_K^2))$，于是 $\mathrm{KL}(q_\phi||p)$ 容易算得（详细推导见附注 1），记为 **KL 正则**：

$$
\begin{aligned}
\mathcal{L}_{KL} &= \frac{1}{2}\left(-\log|\Sigma| + \mathrm{tr}(\Sigma) + \mu^T\mu - k\right) \\
&\xlongequal{\Sigma=\mathrm{diag}(\sigma_1^2,\cdots,\sigma_K^2)} -\sum_{i=1}^{k} \log\sigma_i + \frac{1}{2}(\sigma^T\sigma + \mu^T\mu - k)
\end{aligned} \tag{2.2}
$$

其中 $\sigma = \left(\sigma_1^2, \cdots, \sigma_K^2\right)^T$.

第二项，对编码器参数的更新：这一个对重建图像和原始图像差距的二范数度量，记为**图像重建损失**：

$$
\mathcal{L}_{img} := ||x - p_\theta(x|z)||_2^2 \tag{2.3}
$$

注：此处的解码器输出应该为 $\mu_\theta(z)$，但为了不与编码器中的 μ 混淆，我们都将其记为 $p_\theta(x|z)$.

式2.2,2.3就是 VAE 的全部损失，到目前为止都与类别无关，下面我们将借用 VAE 的隐变量和 ISDA 的损失函数，对隐参数加入类别信息.

2.1.1 ISDA 损失向特征中加入类别信息

值得注意的是 $\mathrm{KL}(q_\phi||p)$ 是一个只对 q_ϕ 的均值和方差的范数进行了限制，而对方差的方向没有任何约束，我们期望在隐参数的最大方差变化方向是**保持类别相同的前提下**的一个隐变量变化，也就是说，如果我们得到了 $q_\phi(z|x) = \mathcal{N}\left(\mu, \mathrm{diag}(\sigma_1^2, \cdots, \sigma_K^2)\right)$ 这样的一个分布，如果我们按照 σ_i 从小到大排序得到 $\sigma_{r_1} \geqslant \sigma_{r_2} \geqslant \cdots \geqslant \sigma_{r_K}$，那么如果我们将 μ 向 σ_{r_1} 方向进行平移得到 $\mu \pm \lambda\sigma_{r_1}$（其中 λ 为步长），那么重建结果 $p_\theta(x|\mu \pm \lambda\sigma_{r_1})$ 就是当前图像 x 在第 r_1 个特征上进行的一个数据增强结果，且应该**保持类别和** x **相同**.

如何加入这个类别这个先验信息呢？我们考虑构造一个特殊的分类损失函数（ISDA 损失），我们假设当前样本为 (x, y)，对应的隐参数分布为 $q_\phi(z) = \mathcal{N}(\mu, \Sigma)$，考虑在输出 μ, Σ 后通过一个全链接层 $f_{w,b}(x) = w^Tx + b : \mathbb{R}^K \to \mathbb{R}^C$ 直接得到对类别的预测值，其中 K 为隐空间维度，C 为总类别数目，可参考图1.

设 $\tilde{z} \sim N(\mu, \Sigma)$ 为当前隐分布上的采样，其都能在最终的交叉熵损失中被最小化，即

$$
\begin{aligned}
&\min_{w,b,\Sigma} \mathbb{E}_{\tilde{z}\sim N(\mu,\Sigma)}[-\log \operatorname{softmax}(w^T\tilde{z}+b)_y] \\
&= \mathbb{E}_{\tilde{z}\sim N(\mu,\Sigma)}\left[\log \sum_{j=1}^{C} \exp\left\{(w_j^T - w_y)\tilde{z} + b_j - b_y\right\}\right] \\
(Jensen's\ inequality) &\leqslant \log \mathbb{E}_{\tilde{z}\sim N(\mu,\Sigma)} \exp\left\{(w_j^T - w_y^T)\tilde{z} + b_j - b_y\right\} \\
&= \log \sum_{j=1}^{C} \exp\left\{\frac{1}{2}(w_j - w_y)^T\Sigma(w_j - w_y) + (w_j - w_y)^T\mu + b_j - b_y\right\} \\
&\xlongequal{iid} \log \sum_{i=1}^{C} \exp\left\{\frac{1}{2}\left((w_j - w_y)^2\right)^T\sigma^2 + (w_j - w_y)^T\mu + b_j - b_y\right\} \\
&=: \mathcal{L}_{class}
\end{aligned}
\tag{2.4}
$$

其中第四行的详细步骤请见附注 2，通过这一损失就能解决分类问题和分布的方差方向问题.

综上，我们的模型总共包含三个损失：

$$
\mathcal{L}(\phi,\theta,w,b) = c_1\mathcal{L}_{KL}(\phi) + c_2\mathcal{L}_{img}(\theta) + \mathcal{L}_{class}(\phi,w,b) \tag{2.5}
$$

其中 $c_1 = 2.5\times 10^{-3}$ 为 KL 正则的加权系数，$c_2 = 20$ 为图像损失的加权系数.

3 Experiment

代码：KataCV/G-VAE，全部代码的功能及解释在 GitHub 上都有详细解释.

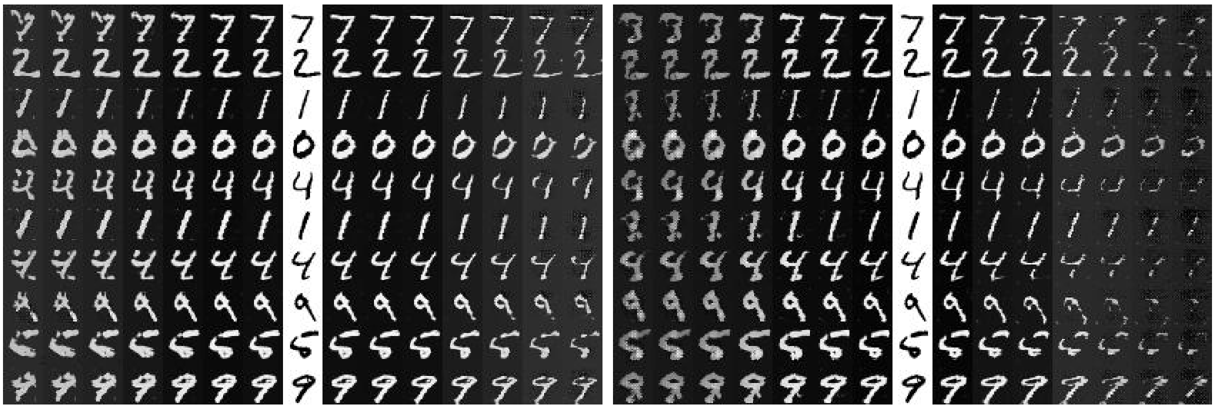
训练细节：

1. 由于 $\sigma \geqslant 0$，且公式中出现的均为 σ^2 形式，所以模型的直接输出结果为 $\log\sigma^2$.
2. ISDA 损失中存在有多个 exp 求和后求 log，存在溢出的可能，所以需要分步计算 $\log(\mathrm{e}^{x_1}+\mathrm{e}^{x_2})$，从而避免精度溢出，需要用到 `jax.lax.scan` 实现高效循环.
3. KL 损失作为正则项其系数不能过小，否则方差的分布会接近均匀分布，导致重建后的图像具有大量的空缺像素，取 $c_1 = 2.5\times 10^{-3}$.
4. 由于该网络为多任务学习，对于不同损失的梯度尺度大小并不相同，当前有些算法能够对梯度进行标准化（通过动态调整不同损失前的加权系数），但是较为复杂，我们直接通过人工调试找到了较为稳定系数 $c_2 = 20$.
5. 图像损失使用 ℓ^2 比 ℓ^1 要好，收敛速度更快，生成的图像能够更连续.（和变分结果中的正态分布对应）
6. 由于直接使用图像方差 σ^2 作为特征变化方向会导致特征变化不明显，所以我们引入一个方差最大裁剪比例 λ_{clip}，将整体方差从大到小排序后，只保留前 λ_{clip} 的方差，其余方差均置为 0.

3.1 MNIST

在 MNIST 上的数据生成结果如下，其中中间的白色部分为原始图像，设其编码特征为 z，则左右第一列通过 z 重建后的图像，继续向左和向右的第 i 列的图像为分别为

$z \pm 0.5\sigma \cdot i$ 的特征重建得到. 下图分别为 G-VAE 和 VAE 重建后的结果（3σ 范围），可以看出来 G-VAE 能够保持重建出的图像类别上的一致性. 方差裁剪系数 $\lambda_{clip} = 0.1$，特征空间维度为 512.

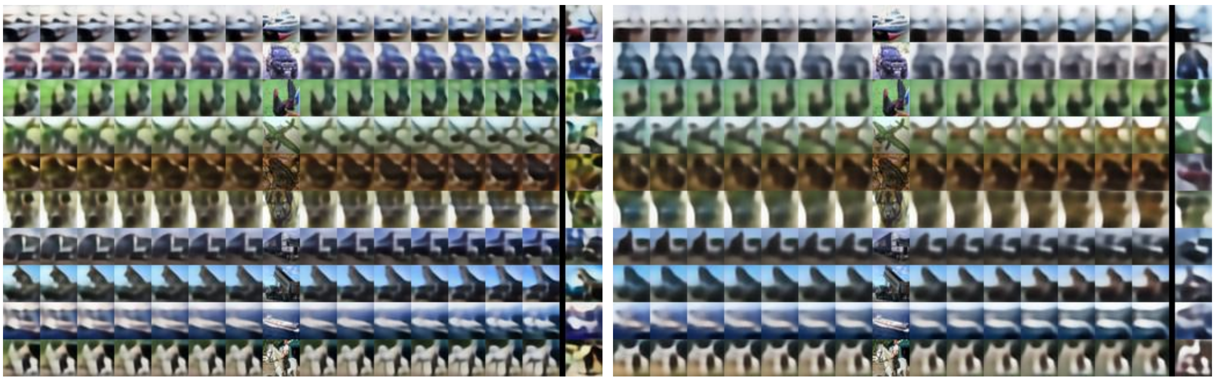


(a) G-VAE 对 MNIST 进行数据增强　(b) VAE 对 MNIST 进行数据增强

图 2: G-VAE 和 VAE 对 MNIST 数据进行增强，$\lambda_{clip} = 0.1, K = 512$

3.2 cifar10

在 cifar10 上的数据生成结果如下，生成方法和 MNIST 一致，不同的是最右列为直接加入 Gauss 噪声后的增强结果. 可以看出 G-VAE 相对于 VAE 的生成结果，清晰度相对更高，能够保持当前类别的特征.（例如第二行是对车辆的颜色进行的变化，第七行是对火车的车厢高度进行的变化，第九行是对船中烟囱的变化）方差裁剪系数 $\lambda_{clip} = 0.1$，特征空间维度为 2048.

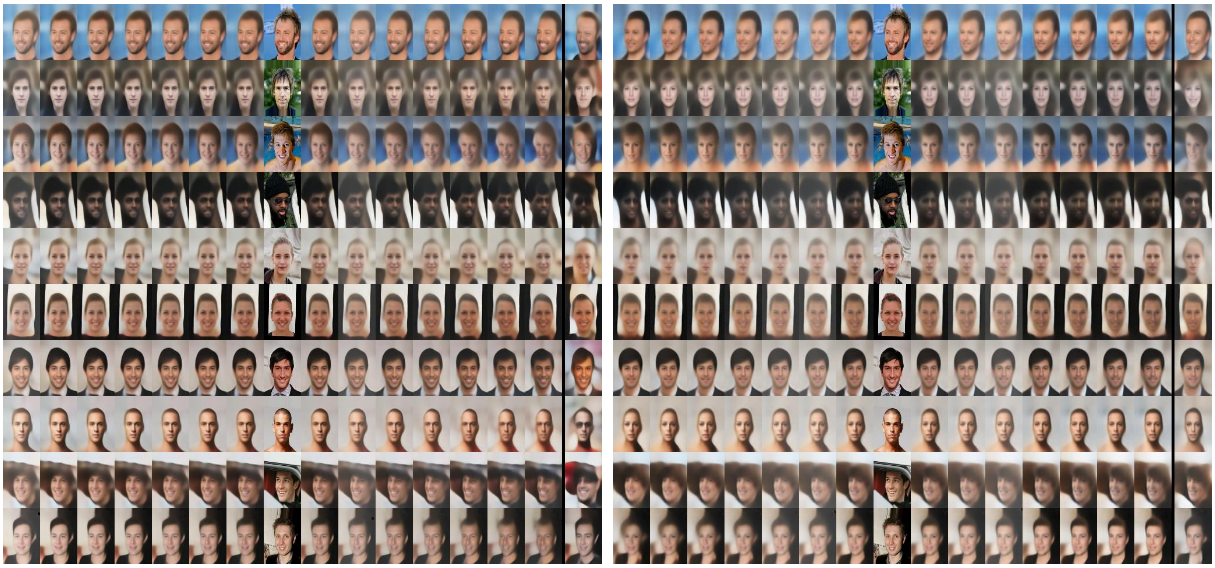


(a) G-VAE 对 cifar10 进行数据增强　(b) VAE 对 cifar10 进行数据增强

图 3: G-VAE 和 VAE 对 cifar10 数据进行增强，$\lambda_{clip} = 0.1, K = 2048$

3.3 Celeba

在人脸数据集 Celeba 上生成数据如下，生成方法和 MNIST 一致，将方差裁剪系数 $\lambda_{clip} = 0.1$ 改为 0.05，特征空间维度为 2048. 我们将图像的标签按照性别和是否微笑划分为了 4 个类别进行分类.



(a) G-VAE 对 Celeba 进行数据增强 (b) VAE 对 Celeba 进行数据增强

图 4: G-VAE 和 VAE 对 Celeba 数据进行增强，$\lambda_{clip} = 0.05, K = 2048$

我们还将同一幅图沿着不同的特征方向进行了变化：

(a) 不同肤色和性别上的变化 (b) 不同角度上的变化

图 5: 上图中所有的图片均为 G-VAE 从一个图片的特征，在不同的类别对应的方差均值下生成得到，类别上只有微笑和性别两个分类指标，但是图像的渐变中还体现出旋转的特征变化

最后我们还尝试使用 PCA 对特征进行降维可视化，如下图所示：

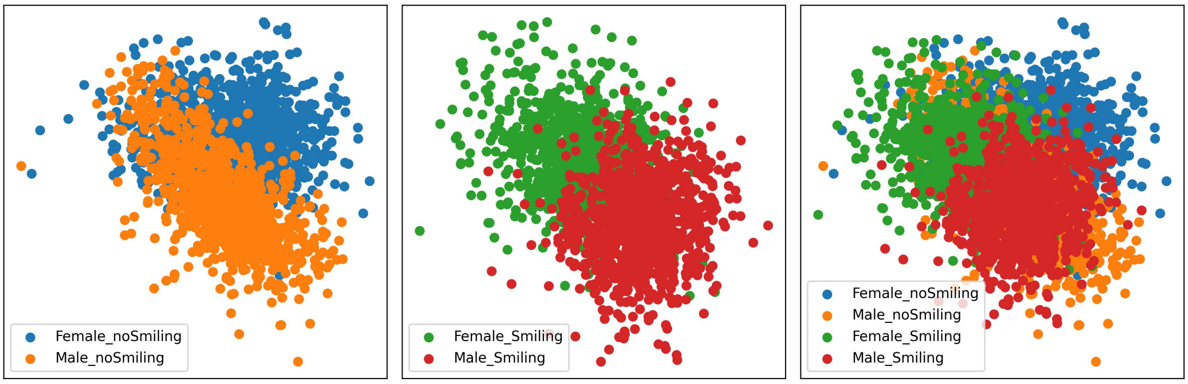


图 6: 上图从 Celeba 的每个类别中分别采样了 1000 张图片，然后利用 PCA 将特征维度从 2048 降维至 2 维，左边两幅是以性别作为绘图的划分标准，可以看出 G-VAE 确实可以有效地其进行划分，最右侧的图片是将 4 个类别同时绘制出来的结果.

4 附注 1（KL 散度计算）

设 $q(x) \sim N(\mu, \Sigma)$ 其中 $\Sigma = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_K^2)$（即 $X_1, \cdots, X_K$ 独立同分布）为 K 维正态分布，$p(x) \sim N(0, I_K)$ 为 K 维标准正态分布，则

$$
\begin{aligned}
&\mathrm{KL}(q||p) \\
&= \int \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\} \log\left\{\frac{\frac{1}{\sqrt{(2\pi)^k|\Sigma|}}\exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}}{\frac{1}{\sqrt{(2\pi)^k}}\exp\left\{-\frac{1}{2}x^Tx\right\}}\right\} \mathrm{d}x \\
&= \frac{1}{2}\left\{-\log|\Sigma| - \int \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\} \left(\color{red}{x^Tx} - \color{blue}{(x-\mu)^T\Sigma^{-1}(x-\mu)}\right) \mathrm{d}x\right\}
\end{aligned}
\tag{4.1}
$$

分别考虑上式中的第二项（红色）和第三项（蓝色）:

$$
\text{第二项} = \mathbb{E}_X(X^TX) = \mathbb{E}\left[\sum_{i=1}^K X_i^2\right] \overset{iid}{=\!=} \sum_{i=1}^K \mathbb{E}\left[X_i^2\right] = \sum_{i=1}^K \sigma_i^2 + \mu_i^2 = \mathrm{tr}(\Sigma) + \mu^T\mu
$$

第三项：由于 $(x-\mu)^T\Sigma^{-1}(x-\mu) = x^T\Sigma x - 2x^T\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\mu$，分别计算：

- $\mathbb{E}_X(X^T\Sigma^{-1}X) = \mathbb{E}_X\left[\sum_{i=1}^K \frac{1}{\sigma_i^2}X_i^2\right] = \sum_{i=1}^K \frac{1}{\sigma_i^2}(\sigma_i^2 + \mu_i^2) = K + \sum_{i=1}^K \frac{\mu_i^2}{\sigma_i^2}$
- $\mathbb{E}_X(X^T\Sigma^{-1}\mu) = \sum_{i=1}^K \frac{\mu_i}{\sigma_i^2}\mathbb{E}_X(X_i) = \sum_{i=1}^K \frac{\mu_i^2}{\sigma_i^2}$
- $\mathbb{E}_X(\mu^T\Sigma^{-1}\mu) = \mu^T\Sigma^{-1}\mu = \sum_{i=1}^K \frac{\mu_i^2}{\sigma_i^2}$

$$
\text{第三项} = K + \sum_{i=1}^K \frac{\mu_i^2}{\sigma_i^2} - 2\sum_{i=1}^K \frac{\mu_i^2}{\sigma_i^2} + \sum_{i=1}^K \frac{\mu_i^2}{\sigma_i^2} = K
$$

综上

$$\mathrm{KL}(q||p) = \frac{1}{2}\left(-\log|\Sigma| + \mathrm{tr}(\Sigma) + \mu^T\mu - K\right)$$

4.1 附注 2（简化 ISDA 损失）

计算 $\mathbb{E}_{\tilde{z}\sim N(\mu,\Sigma)}\exp\left\{(w_j^T - w_y^T)\tilde{z} + b_j - b_y\right\}$，不妨令 $w = w_j - w_y, b = b_j - b_y$，则

$$\begin{aligned}
&\mathbb{E}_{x\sim N(\mu,\Sigma)}\exp\{wx+b\}\\
=&\mathbb{E}_{x\sim N(\mu,\Sigma)}\frac{1}{\sqrt{2\pi|\Sigma|}}\exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) + w^Tx + b\right\}\\
=&\mathbb{E}_{x\sim N(\mu,\Sigma)}\frac{1}{\sqrt{2\pi|\Sigma|}}\exp\left\{-\frac{1}{2}\left(x-(\mu+w^T\Sigma w)\right)^T\Sigma^{-1}\left(x-(\mu+w^T\Sigma w)\right) + \frac{1}{2}w^T\Sigma w + w^T\mu + b\right\}\\
=&\exp\left\{\frac{1}{2}w^T\Sigma w + w^T\mu + b\right\}\\
=&\exp\left\{\frac{1}{2}(w_j-w_y)^T\Sigma(w_j-w_y) + (w_j-w_y)^T\mu + b_j - b_y\right\}
\end{aligned} \tag{4.2}$$